A Survey of Job Scheduling in Cloud Compuing Environment

Shyam Sunder Pabboju[#1], Satya Shekar Varma P[*2], Damera Vijay Kumar[#3]

[#1]*Assistant Professor, CSE, MGIT, Hyderabad, India*
[#2]*Assistant Professor, CSE, MGIT, Hyderabad, India*
[#3]*Assistant Professor, IT, MGIT, Hyderabad, India*

Abstract*—Cloud Computing is an emerging technology, recent years several enterprises are shifting to Cloud environment. One of the major challenges of Cloud computing is Job Scheduling. Scheduling is the allocation of jobs on to the resources in particular time. Scheduling is a technique which is used to improve overall execution time of the job. In this paper, we have presented a survey on different job scheduling algorithms in cloud computing environment.*

Keywords *- Cloud Computing, Virtualizatiion, Quality of Service(QoS), Scheduling.*

I. INTRODUCTION

Cloud Computing[1] is a construct which allows user to access applications that actually resides at a place other than your computer or other internet-connected device. The Reason for the Success of the Cloud Computing is by underlying technology called Virtualization[2]. Virtualization is a technology that allows executing two or more operating system side-by-side on just one PC. Ease of use Many applications are having a limited number of concurrent tasks, thus making number of cores idle, these problems can be overcome by using Virtualization, allocating group of cores to an OS that can run it concurrently. It enables the service provider to offer virtual machines for jobs rather than physical server machines. Virtual machines provide flexibility and mobility through easy migration which enables dynamic mapping of VM's to available resources.

Job Scheduling[4] is a key role in cloud computing environment. Job Scheduling is a process of allocating jobs onto available resources in particular time. Job Scheduling is biggest and challenging issue. The main aim of job scheduling is to improve the performance and Quality of Service (QoS)[3] and at the same time maintaining the efficiency and fairness among jobs and reduce the execution cost. Such process has to respect constraints given by the jobs and the cloud. Scheduling should satisfy constraints provided by users and cloud providers. Constraints given by user like giving deadline of completion job or completion job within budget. Constraints given by cloud providers includes maximum resource utilization and maximize the benefit i.e maximum return on investment.

A.Basic terminology

Tasks: represents a minimal computational unit to run on a node.

A Task is considered as an indivisible schedulable unit.

Task could be independent(or loosely coupled) or there could be dependencies.

Jobs: A job is a computational activity made up of several tasks that could require different processing capabilities and could have different resource requirements(CPU, number of nodes, memory, software libraries etc) and constraints, usually expressed within the job description.

Each job may have various parameters such as required data, desired completion time often called deadline, expected execution time, job priority etc.

Resources: A resource something that is required to carry out an operation for example- a processor for processing, a data storage device, or a network link for data transporting.

B.Quality of Service

Every user wants the best Quality of Service(QoS) for its application. Quality of Service is the ability to provide different priority to different jobs and users or to guarantee a certain level of performance to a job. Quality of Service management is the problem of allocating resources to the application to guarantee a service level along dimensions such as performance, availability and reliability.

If the QoS mechanism is supported it allows the users to specify desired performance for their jobs such as

- Completion before the given deadline
- Dedicated access to resources during some time period(advanced reservation)
- Requested amount of resources (CPUs, RAM, HDD, network bandwidth), etc.

Such request is formally established between the user and the resource manager (scheduler) through negotiation which produces a formal description of the guaranteed service called Service Level Agreement(SLA).

Scheduling problem is related to two types of users:

- Cloud Consumers
- Cloud providers

What users want:

Cloud Consumers:

- Execute jobs for solving problem of varying size and complexity.
- Benefit by electing ad aggregating resources wisely
- Tradeoff time and cost

Cloud Providers:

- Contribute ("idle") resources for executing consumer jobs.
- Benefit by maximizing resource utilization.
- Tradeoff local requirement and market chance.

II. FRAMEWORK OF JOB SHEDULING IN CLOUD COMPUTING ENVIRONMENT

SLA Monitor

When a consumer first submits the service request,the SLA Monitor interprets the submitted request for QoS requiremens before determining whether to accept or reject the request.

It is also responsible to monitor the progress of the submitted job.If any validation is observed from SLA it has to act immediately for correction action.(eg resource fails).

Resource Discovery and monitoring

Resource Discovery may be described basically as the task in which the provider should find appropriate resources in order to comply with incoming consumer requests.

Considering one of the key features of cloud computing is the capability of acquiring and releasing resource on demand, resource monitoring should be continuous.

Task Scheduling

The input of task scheduling algorithm is normally as abstract model which defines tasks without specifying the physical location of resources on which the tasks are executed.

Reschedule: When a task can't completed due to processor failure or a disk failure or other problems, the uncompleted tasks could be rescheduled in the next computation.

Scheduling Optimizer

After acquiring information about available resources in the cloud (during the discovery phase),a set of appropriated candidate is highlighted.

The Resource selection mechanism elects the candidate solution that fulfills all requirements and optimizes the usage of the infrastructure.The resource selection may be done using an optimization algorithm such as meta heuristic algorithm like genetic algorithm,Ant colony[4],particle swarm optimization(pso) for cloud.

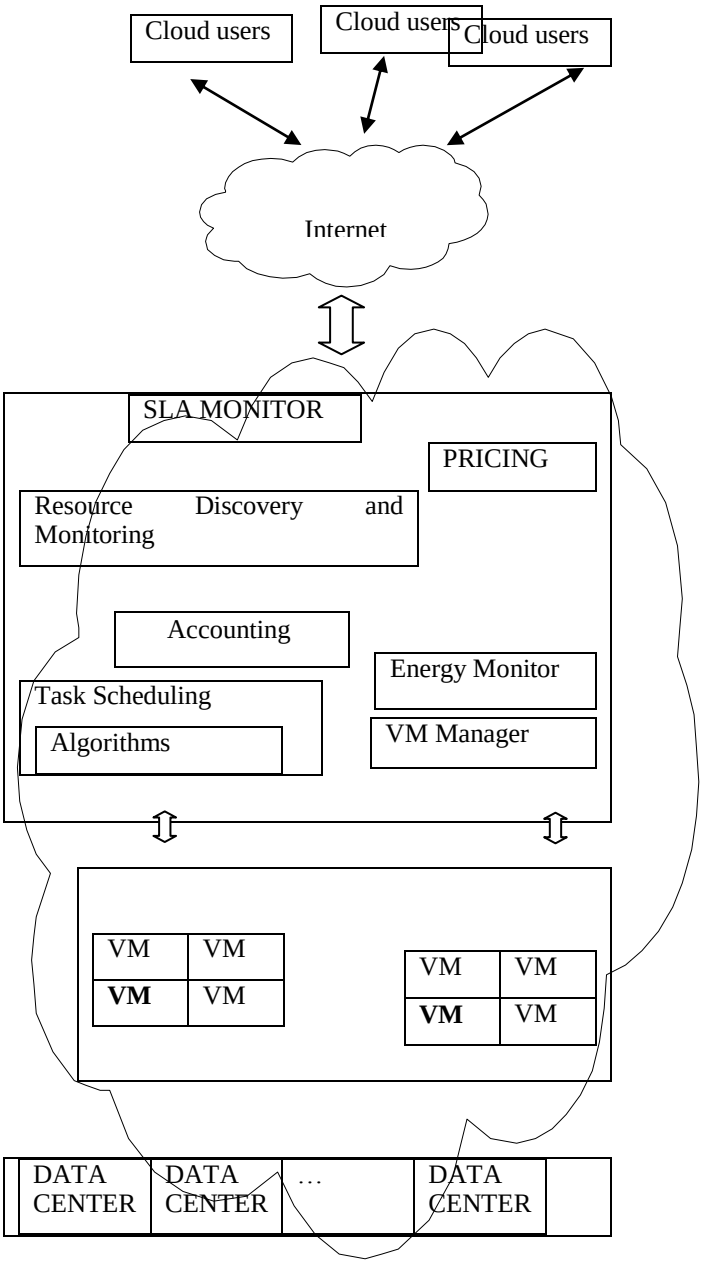

Fig.1 Framework for job scheduling in cloud computing environment

Advance Resource Reservation Monitor

Provides QoS guarantees for accessing various resource across data centers.

Users are able to secure resources acquired in the future which are important to ensure successful completion of time critical applications such as real time and workflow applications (or) parallel

applications aquring a number of processors to run.

Provider can predict future demand and usage more accurately.

Data centers & Virtual machine

Data centers are physical machine severs,which are not used directly by users. Data centers are converted into virtual machine by virtualization concept. Users jobs will run on these virtual machines through scheduling.

VM Manager

VM Manager manages all the virtual Machines.

III. DIFFERENT TYEPS OF SCHEDULING

The different types of job scheduling[5] in cloud computing environment are:

- Static Vs dynamic Scheduling
- Centralized Vs Hierarchical Vs Distributed Scheduling
- Preemptive Vs NonPreemptive Scheduling
- Immediate Vs Batch Mode Scheduling
- Independent Vs workflow Scheduling

Static Vs Dynamic Scheduling

Static Scheduling

Information regarding all resources as well as all tasks in the application to be available by the application is scheduled.

No job failure and resources are assumed available all the time.

Dynamic Scheduling

Jobs are dynamically[6] available for scheduling over time by the scheduler with no issues, to be able of determining run time in advance.

The dynamics of job execution, which refers to the situation when job execution could fail due to some resources failure or job execution could be stopped due to the arrival in the system of high priority jobs when the case of preemptive mode is considered.

The dynamics of resources, in which workload of resources can significantly vary over time.

Centralized Vs Hierarchical Vs Distributed Scheduling

Centralized Scheduling

- Centralized and decentralized scheduling differ in the control of resources and knowledge of the overall system.
- In case of centralized scheduling, there is more control on resources as the scheduler has knowledge of the system by monitoring of the resource state.
- Advantages: ease of implementation, efficiency and more control and monitoring of the resource state.
- Disadvantage: lacks scalability, fault tolerance and efficient performance.

Hierarchical Scheduling

Allow one to coordinate different scheduler at a certain level.

Schedulers at the lowest level in the hierarchy have the knowledge of the resources

Disadvantages: lack of scalability and fault tolerance, yet it scales better and is more fault tolerant than centralized schedulers.

Decentralized or Distributed Scheduling

No central entity controlling the resources.Scheduling decisions are shared by multiple distributed schedulers.Less efficiency than centralized scheduler.

Immediate Vs Batch Mode Scheduling

Immediate/online mode

Immediate/online mode in which scheduler schedules any recently arriving job as it arrives with no waiting time for next time interval on available resources at that time.

Batch/Offline Mode Scheduling

The scheduler holds arriving jobs as groups of problems to be solved over successive time intervals, so that it is better to map a job for suitable resources depending on its characterstics.

Independent Vs workflow/Dependent Scheduling

Independent scheduling

Tasks can be run independently.

Workflow scheduling

- Tasks are dependent on each other
- Dependency means as there are precedence orders existing in tasks ,that is a task cannot start until all its parent are done.
- Workflows represented by Directed Acyclic Graph(DAG).Each task can start its execution when all preceding task in DAG are already finished.

IV.SCHEDULING INDEPENDENT TASKS

A.Min-Min Heuristic Algorithm[7]

- For each task determine its minimum completion time over all machines.
- Over all tasks find the minimum completion time.

- Assign the task to the machine that gives this completion time.
- Iterate till all the tasks are scheduled.

Example of Min-Min

	T1	T2	T3
M1	140	20	60
M2	100	100	70

	T1	T3
M1	160	80
M2	100	70

Stage 1: Stage 2: Stage 3:

T1-M2=100 T1-M2-100 T1-m1=160

	T1
M1	160
M2	170

T2-M1=20 T3-m2=70

T3-M1=60

Assign T2 to M1 Assign T3 to M2 Assign T1 to M1

M1	T2	T1
	20	160

M2	T3
	80

B.Max-Min Heuristic Algorithm

- For each task determine its minimum completion time over all machines.
- Over all tasks find the maximum completion time.
- Assign the task to the machine that gives this completion time.
- Iterate till all the tasks are scheduled.

Example of Min-Min

	T1	T2	T3
M1	140	20	60
M2	100	100	70

	T2	T3
M1	80	60
M2	200	170

	T3
M1	80
M2	200

Stage 1: Stage 2: Stage 3:

T1-M2=100 T1-M2-100 T1-m1=160

T2-M1=20 T3-m2=70

T3-M1=60

Assign T2 to M1 Assign T3 to M2 Assign T1 to M1

Gantt Chart:

M1	T3	T2
	60	80

M2	*T1*
	100

C. Sufferage Heuristic Algorithm

- For each task determine the difference between its maximum and second minimum completion time over all machines (sufferage).
- Over all tasks find the maximum sufferage.
- Assign the task to the machine that has reached in obtaining minimum completion time.
- Iterate till all the tasks are scheduled.

	T1	T2	T3
M1	140	20	60
M2	100	100	70

	T1	T3
M1	160	80
M2	100	70

	T3
M1	80
M2	170

Stage 1: Stage 2: Stage 3:

T1=40 T1=60 T3=90

T2=80 T3=10

T3=10

Gantt Chart:

M1	T2 20	T3 80
M2	T1 100	

D. First Come First Serve Algorithm

This algorithm is simplest and fastest. In this algorithm Jobs are as first come first serve basis as like queue data strucuture.

E. Round Robin algorithm

In this scheduling algorithm[9], each job given a limited amount of time called a time-slice or a quantum in FIFO manner. If a job does not complete execution before its time expires, the CPU is pre-empted and given to the next process waiting in a queue. And the preempted job is placed at the end of the ready queue and processed in the next time slice or quantum.

F. Most fit task scheduling algorithm

In this the job which fit best is allocated first.

G. Priority scheduling algorithm[7]

Each process is assigned a priority, and then based on priority processes are allowed to be executed. Equal priority processes are executed in FCFS order.

V. WOKFLOW/DEPENDENT SHEDULING ALGORITHMS

Tasks composing a job have the precedence orders[8].The Directed Acyclic Graps(DAGs) are represented to use these tasks.In DAGs , a node represents a task and directed edge denotes the precedence orders between the tasks.In some cases .weights can be added to nodes and edges to express computational costs and communicating costs respectively.

A. List Scheduling Heuristics Algorithm

An orderd list of taks is constructed by assigning priority to each task.Tasks are selected on priority order and scheduled in order to minimize a predefined cost function.

B. Clustering Heuristics

In Clustering Heuristics algorithm, the group of tasks together formed into cluster.Tasks in the same cluster are
scheduled on the same resource.

VI.CONCLUSION

Efficient scheduling algorithm always plays important role in cloud computing. Scheduling is a major issue in cloud computing. In this paper we surveyed various existing scheduling algorithms related to Cloud Scheduling. In Cloud computing main goal of job scheduling is to maximize utilization of resources and satisfy the user requirement. The existing algorithm is related to the parameters like makespan, completion time, cost and performance. In existing algorithm when priority is considered starvation problem is created. So, there are many aspects of research based on priority and improving parameters like makespan, completion time, Performance, Average Utilization Ratio etc.

VII.REFERENCES

[1] Anthony T. Velte, Toby J. Velte, Robert Elsenpeter, “Cloud Computing, A Practical Approach.

[2] Aasys, Virtualization Basics, Vol. 6, Issue 9, September 2008.

[3] Abdelzahir Abdelmaboud, Dayang N.A. Jawawia, Imran Ghania, Abubakar Elsafia, Barbara Kitchenhamb : Quality of service approaches in cloud computing: A systematic mapping study, Journal of Systems and Software (Science Direct) , Volume 101, March 2015, Pages 159–179.

[4] Umarani Srikanth G., V. Uma Maheswari, P. Shanthi, Arul Siromoney, "Tasks Scheduling using Ant Colony Optimization", Journal of computer Science 8 (8) pp 1314-1320, Science Publications(2012).

[5] R. Bajaj and D. P. Agrawal, “Improving Scheduling of Tasks in a Heterogeneous Environments,” IEEE Transactions on Parallel and Distributed Systems, Vol.15,No. 2, 2004, pp. 107-118 2004.

[6] J.Li, M.Qiu, X.Qin, “Feedback Dynamic Algorithms for Preemptable Job Scheduling in Cloud System”, IEEE, 2010,used dynamic min-min algorithm.

[7] Huankai Chen, Frank Wang, Na Helian, Gbola Akanmu “User- Priority Guided min-Min Scheduling Algorithm for Load Balancing in cloud computing” National conference on Parallel computing technologies , pp 1-8, 21-23 Feb 2013(IEEE).

[8] Shamsollah Ghanbari, and Mohamed Othman, “A Priority based Job Scheduling Algorithm in Cloud Computing”, International Conference on Advances Science and Contemporary Engineering, 2012.

[9] S. Xavier and S. J. Lovesum, "A survey of various workflow scheduling algorithms in cloud environment," International Journal of Scientific and Research Publications, 3(2), 2013